

20 May 2013

Dr. Harold Hawkins
ONR Code 341
Office of Naval Research
875 North Randolph SL
Arlington. VA 22203-1995

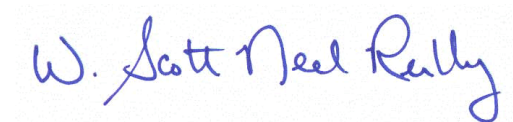
Reference: US Navy Contract N00014-12-C-0653: "The Model Analyst's Toolkit: Scientific Model Development, Analysis, and Validation"
Charles River Analytics Contract No. C12186

Subject: Contractor's Quarterly Status Report #2
Reporting Period: 20-February-2013 to 19-May-2013

Dear Dr. Hawkins,

Please find enclosed 1 copy of the Quarterly Status Report for the referenced contract. Please feel free to contact me with any questions regarding this report or the status of the "The Model Analyst's Toolkit: Scientific Model Development, Analysis, and Validation" effort.

Sincerely,



W. Scott Neal Reilly
Principal Investigator

cc: Michael Hession, DCMA
Annetta Burger, ONR
Whitney McCoy, Charles River Analytics

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE MAY 2013		2. REPORT TYPE		3. DATES COVERED 00-02-2013 to 00-05-2013	
4. TITLE AND SUBTITLE The Model Analyst's Toolkit: Scientific Model Development, Analysis, and Validation				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Charles River Analytics, 625 Mount Auburn Street, Cambridge, MA, 02138				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 21	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

Charles River Analytics

Monthly Technical Progress Report No. R12186-03
Reporting Period: February 20, 2013 to May 19, 2013
Government Contract No. N00014-12-C-0653
Charles River Analytics Contract No. C12186

The Model Analyst's Toolkit: Scientific Model Development, Analysis, and Validation

Quarterly Status Report

Principal Investigator: Scott Neal Reilly

Charles River Analytics
625 Mount Auburn Street
Cambridge, MA 02138
617-491-3474

May 20, 2013

Distribution Statement A: Approved for public release; distribution is unlimited.

The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Agency position, policy, or decision, unless so designated by other official documentation.

1. Executive Summary

The proposed research effort builds on and extends the work of the previous ONR-funded “Validation Coverage Toolkit for HSCB Models” project. The overall objectives of the on-going research program are:

- Help scientists create, analyze, refine, and validate rich scientific models
- Help computational scientists verify the correctness of their implementations of those models
- Help users of scientific models, including decision makers within the US Navy, to use those models correctly and with confidence
- Use a combination of human-driven data visualization and analysis, automated data analysis, and machine learning to leverage human expertise in model building with automated analyses of complex models against large datasets

Specific objectives for the current effort include:

- **Fluid temporal correlation analysis.** Our objective is to design a new method for performing temporally fluid correlation analysis for temporal sets of data and implement the method as a new prototype component within the Model Analyst’s Toolkit (MAT) software application.
- **Automated suggestions for model construction and refinement.** Our objective is to design and implement a prototype mechanism that learns from data how factors interact in non-trivial ways in scientific models.
- **Data validation and repair.** Our objective is to design and implement a prototype capability to identify likely errors in data based on anomalies relative to historic data and to use models of historic data to offer suggested repairs.
- **System prototyping.** Our objective is to incorporate all improvements into the MAT software application and make the resulting application available to the government and academic research community for use in scientific modeling projects.
- **Evaluation of applicability to multiple scientific domains.** Our objective is to ensure (and demonstrate) that MAT can be applied to a wide range of scientific domains by identifying and building at least one neurological and/or physiological model and analyze the associated data with MAT, making any extensions to the MAT tool that are needed to support the analysis of such a model.

2. Overview of Problem and Technical Approach

2.1. Summary of the Problem

One of the most powerful things scientists can do is to create models that describe the world around us. Models help scientists organize their theories and suggest additional experiments to run. Validated models also help others in more practical applications. For instance, in the hands of military decision makers, human social cultural behavior (HSCB) models can help predict instability and the socio-political effects of missions,

whereas models of the human brain and mind can help educators and trainers create curricula that more effectively improve the knowledge, skills, and abilities of their pupils.

While there are various software tools that are used by the scientific community to help them develop and analyze their models (e.g., Excel, R, Simulink, Matlab), they are largely so general in purpose (e.g., Excel, R) or so focused on computational models in particular (e.g., Simulink, Matlab), that they are not ideal for rapid model exploration or for use by non-computational scientists. They also largely ignore the problem of validating the models, especially when the models are positing causal claims as most interesting scientific models do. To address this gap, Charles River Analytics undertook the “Validation Coverage Toolkit for HSCB Models” project with ONR. Under this effort, we successfully designed, implemented, informally evaluated, and deployed a tool called the Model Analyst’s Toolkit (MAT), which focused on supporting social scientists to visualize and explore data, develop causal models, and validate those models against available data (Neal Reilly, 2010; Neal Reilly, Pfeffer, & Barnett, 2010; Neal Reilly et al., 2011).

As part of the development of the MAT tool, we identified four important extensions to that research program that would further support the scientific modeling process:

- Correlation analyses are still the standard way of identifying relationships between factors in a model, but correlations are fundamentally flawed as a tool for analyzing potentially causal or predictive relationships as they assume instantaneous effects. Even performing correlation analyses with a temporal offsets between streams of data is insufficient as the temporal gap between the causal or predictive event and the following event may not be the same every time (either because of variability in the system being modeled or because of variability introduced by a fixed sampling rate). What we need is a novel way of evaluating the true predictive power across streams of data that can deal with fluid offsets between changes in one stream of data and follow events in the other stream of data.
- Modeling complex phenomena is a fundamentally difficult task. Human intuition and analysis is by far the most effective way of performing this task, but even humans can be overwhelmed by the complexity of modeling the systems they are studying (e.g., socio-political system, human neurophysiology). Automated tools, while not especially good at generating reasonable scientific hypotheses, *are* extremely good at processing large amounts of data. We believe there is an opportunity for computational systems to enhance human scientific inquiry. Under the “Validation Coverage Toolkit for HSCB Models” project, we demonstrated how automated tools could help human scientists to analyze and validate their models against data. We believe a similar approach can be used to help suggest modifications to the human-built models to make them better match the available data. To be useful, however, such automated analyses will need to be rich enough to suggest subtle data interactions that are most likely to be missed by the human scientist. For instance, correlations (especially correlations that take into account fluid temporal displacements) could be used to identify likely

relationships between streams of data, but such an approach would miss complex, non-linear relationships between interrelated factors that cannot be effectively analyzed with simple two-way correlations. For instance, if crime waves are associated with increases in unemployment *or* drops in the police presence, that would be hard to identify with a correlation analysis. We need richer automated data analysis techniques that can extract complex, non-linear, multi-variable relationships between data if we are to effectively suggest model improvements to human scientists.

- Even if a scientific model is sound, if the data sets provided as inputs to the model are unreliable, the results of the model are still suspect. And, unfortunately, data will often be wrong. For instance, HSCB surveys are notoriously unreliable and biased for a variety of reasons, and neurological and physiological data can be corrupted by broken or improperly used sensors. If it were possible to identify when data was unreliable and, ideally, even repair the data, then the models that are using the data could once again be effectively used.
- The MAT tool we developed under the “Validation Coverage Toolkit for HSCB Models” project was focused primarily on assisting social scientists in the analysis, refinement, and validation of HSCB models. In parallel with that effort, however, we also took an opportunity to apply MAT to evaluating neurological and physiological data under the DARPA-funded CRANIUM (Cognitive Readiness Agents for Neural Imaging and Understanding Models) program. We discovered the generality of the MAT tool makes it potentially applicable to a great number of different scientific domains. MAT proved to be a useful, but peripheral tool, in CRANIUM. We believe MAT could be applied to a broader suite of scientific modeling problems than it has been so far.

2.2. Summary of our Approach

To address these identified gaps and opportunities, we are extending MAT’s support for model development, analysis, refinement, and validation; enhancing MAT to analyze and repair data; and demonstrating MATs usefulness in additional scientific modeling domains. Our approach encompasses the following four areas, which correspond to the four gaps/opportunities identified in the previous section:

- **Temporally Fluid Correlation Analysis.** We are designing a new method to perform Temporally Fluid Correlational Analysis on temporal sets of data, and we are implementing the method as a new component within the MAT software application. The version of MAT at the beginning of the new effort supported correlation analysis for temporally offset data; it shifts the two data streams being compared by a fixed offset that is based on the sampling rate of the data (i.e., data that is sampled annually will be shifted by one year at a time), performs a standard correlation on the shifted data, plots the correlation value against the amount of the offset, and then repeats the process for the next offset amount. If two data streams are shifted by a fixed offset (e.g., changes in one stream are always followed by a comparable value in the other stream after a fixed time), then this method will find that offset. Under the current effort, we are expanding

on this capability to support fluid temporal shifts within the data streams. That is, we are making it possible to identify when the temporal offset between the change in the first data stream and its effect in the second stream is not a static amount of time.

- **Automated suggestions for model construction and refinement.** We are designing and implementing a mechanism to learn how factors interact in non-trivial ways in scientific models. In particular, we are developing a method for learning disjuncts, conjuncts, and negations. This mechanism starts with the model developed by the scientist user and make recommendations for possible adjustments to make it more complete by performing statistical data mining and machine learning.
- **Data validation and repair.** Recognizing that data contains errors is plausible once we understand the relationships between data sets. That is, if we are able to develop models of the correlations between sets of data, then we can build systems that notice when these correlations do not hold in new data, indicating possible errors in data. For instance, if we know that public sentiment tends to vary similarly between nearby towns, then when one town shows anomalous behavior, we can reasonably suspect problems with the data. There might be local issues that cause the anomaly, but it is, at least, worth noting and bringing to the attention of the user of the data and model. As MAT is designed to help analyze models and recognize inter-data relationships, it is primed to perform exactly this analysis. Existing methods perform similar types of analysis for environmental data (Dereszynski & Dietterich, 2007; Dereszynski & Dietterich, 2011). For instance, a broken thermometer can be identified and the data from it even estimated by looking at the temperature readings of nearby thermometers, which will generally be highly correlated.
- **Application to multiple scientific modeling domains.** To ensure (and demonstrate) that MAT can be applied to a wide range of scientific domains, we are identifying and building at least one neurological and/or physiological model and analyzing the associated data with MAT, making any extensions to the MAT tool that are needed to support the analysis of such a model. The initial MAT effort focused on HSCB models; by focusing this effort on harder-science models at much shorter time durations, we believe we can effectively evaluate an interesting range of applications of the MAT tool.

3. Current Activities and Status

During the reporting period we made progress on developing methods for temporally fluid correlation analysis (and, more generally, causality analysis techniques), methods for automatic model construction and refinements, and began investigating applications to neurophysiological data. We describe our progress during the current reporting period for each of these three areas in turn.

3.1. Temporally Fluid Correlation Analysis & Causality Analysis

Under the previous research program, we developed the ability to do correlation analyses of temporally offset data. The idea is that if causes precede effects, then we might sometimes find cases where the correlation of two data series where the purported cause is shifted forward in time would have a higher correlation than the two data series unshifted. The initial implementation, however, only worked for static offsets and we expect that the time between the cause and effect might vary slightly between instances of causes and effects. To address these cases, we are exploring more advanced techniques for analyzing such non-statically offset data. The first approach we discuss is called Dynamic Time Warping and is inspired by work in gait recognition where the same gait needs to be recognized even when the subject is slowing down or speeding up. We have also begun exploring two other methods for this type of analysis and describe some initial experiments regarding these techniques in this section.

3.1.1. *Dynamic Time Warping*

Currently, MAT supports the ability to compare two time series data sets and plot their correlation as a function of the phase shift between them. This allows users to compare data sets in which a spike in one results in a spike in the other at a delayed time. The current feature set of MAT, however, does not support comparing data sets which have varying delays, time scaling difference, and varying in sampling rate. The incorporation and utilization of the Dynamic Time Warp Algorithm (DTW) hopes to address such areas and continue to expand the feature set and usability of MAT.

The Dynamic Time Warping Algorithm

Dynamic Time Warping (DTW) is a well established algorithm used to compare two sets of data which may vary in time or speed. The classical use case for DTW is in gait analysis to compare walking versus running, but the algorithm has been used in a wide variety of other contexts such as video and audio processing.

DTW utilizes a cost matrix to determine the optimal “warping” between two data sets defined as the mapping between data points in one series to another that minimizes the sum of distances between the mapped points. One major advantage of DTW compared to other algorithms used for time series comparison is that it can account for missing data points as well as compare two data sets that utilize different time scales or sampling frequencies by compressing or expanding certain areas of the data set.

One limitation of the DTW algorithm is the efficiency in both time and space used by the algorithm. Because it compares all points against one another in a cost matrix, the algorithm runs in quadratic time and uses quadratic space in relation to the size of the input data set. When comparing significantly large data sets, this cost can make the algorithm unusable. However, academic research into optimizations of DTW has resulted in a breakthrough to reduce this cost, providing a linear time approximation algorithm (FastDTW) that performs far better than previous approximations (Chan & Salvador, 2007).

Dynamic Time Warping in MAT

To incorporate the DTW algorithm quickly into MAT, we have utilized the open source and free to use Java library included with the paper describing the FastDTW algorithm (Chan et al., 2007). Both the regular DTW and FastDTW algorithms are implemented in the library and have been incorporated into MAT.

Because we envision the use of the DTW algorithm in MAT to be used more than to compare two time series but also for analysis of causal relationships, we have slightly modified the original DTW algorithm. In order to match time series together, the original DTW algorithm could match points in either direction, allowing some points in the series to shift forward as well as backward to better match up. However, a backward shift in time is nonsensical in this use case and would mislead the user, so we have removed the ability to do so, naming this algorithm ForwardDTW for convenience.

The current functionality of DTW in MAT allows users to select two time series for comparison, displaying the traditional warping lines between the two series. In this case, we are comparing Natural Gas Rents (Green) with Coal Rents (Red).

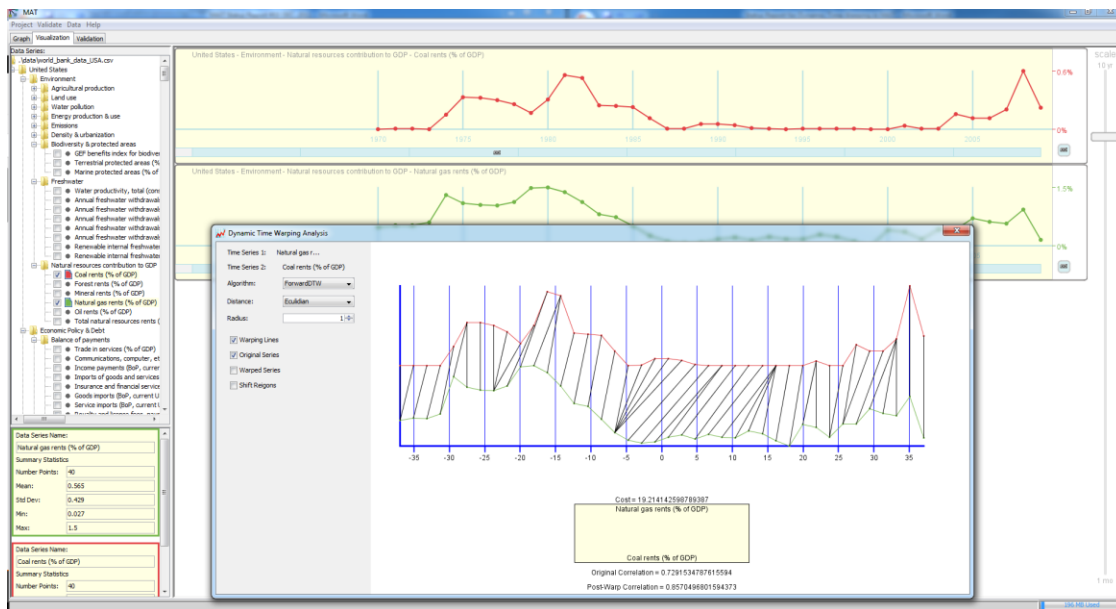


Figure 1. Dynamic Time Warping analysis in MAT

The user can also include the resulting time series if warping process shifts, compresses, and expands the time series to better match the other. In this case, we shift the green line into the brown to better match the red.

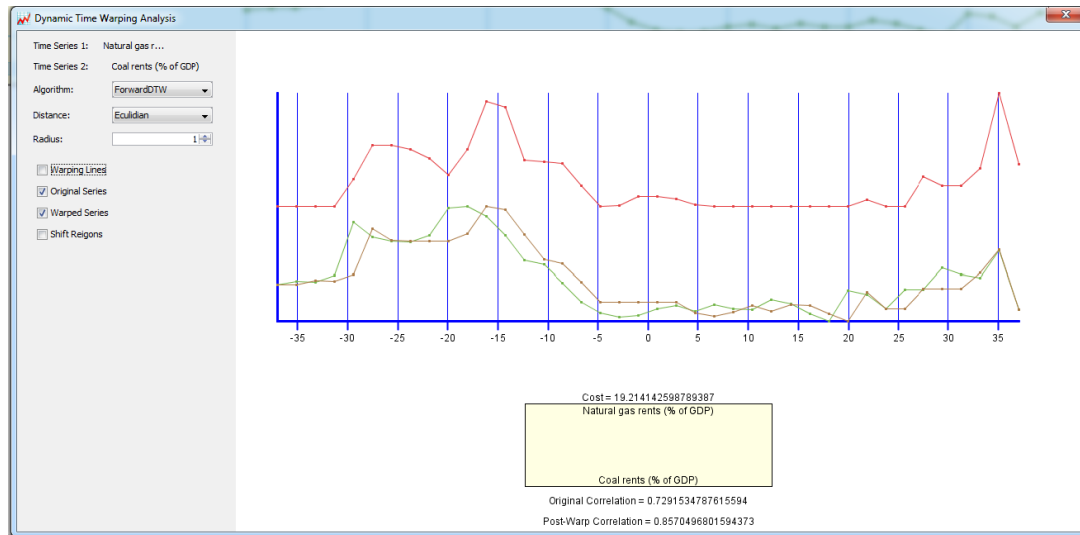


Figure 2. Dynamically timewarped data in Mat

The user can easily hide the lines to better easily view the shift and is given the pre-warp and post-warp correlation values for the two data sets.

One concern with using DTW which we must be aware of is the ability to shift dissimilar data sets without any relationships into those that appear similar. By providing users with some metric or threshold based on the cost of the warp and improvement in correlation, we may be able to prevent such false positives in the majority of cases.

3.1.2. Evaluation of Causality Analysis Techniques

As part of our work in developing a tool for analyzing potentially causally-related time series and for automatically recommending causal models that explain data, we have continued to implement and evaluate complementary approaches to detecting causal relationships between two time series. We compare results obtained by Granger Causality (GC) and Dynamic Time Warping (DTW) approaches for detecting causality, and discuss the importance of selecting a proper representation for the underlying data. A qualitative comparison of GC and DTW methods on World Bank data indicates that both methods are capable of identifying potential causal (or correlative or predictive) relationships between time series. Table 1 illustrates the top 20 World Bank time series likely to share a causal relationship with the Natural Gas Rents records, as determined by DTW cost scores. Table 2 illustrates the same results obtained using GC. Results are shown for comparisons between raw data, scaled data, and scaled and de-trended data series. We note that as GC inherently computes a normalized correlation score, there is no difference between raw and scaled datasets, and we omit GC results on raw data. Results are color-coded based on a series's estimated relevance to the query series: series likely to be correlated are left white; series that may be indirectly correlated are highlighted gray; series deemed unlikely to be causally correlated are highlighted yellow. A graph capturing the causal likelihood score across the top 100 series is illustrated for each column (a lower score indicates a stronger likelihood of correlation/causation). As we

expect few time series to truly exhibit causal relationships, we expect this curve to exhibit an asymptotic slope, in which a small number of time series are closely related to the query data, and similarity to other data rapidly decreases.

A key observation is that DTW is particularly sensitive to significant differences in scale between time series, due to the underlying distance-based matching metric. This can be observed both from the quality of resulting matches, as well as from the shallow shape of the score curve. As a result, datasets must be appropriately normalized prior to performing analysis using DTW. To enable direct comparison between time series using DTW, we normalize datasets to unit standard deviation. A further observation is that certain time series may inherently follow long-term trends that may be present across many time series (i.e., economies tend to grow as a general trend). These trends may conceal relationships between time series that may have significant short-term impacts: to account for this possibility, we additionally apply a linear de-trending process to each time series. This process noticeably impacts the rankings obtained using both DTW and GC techniques; in the case of DTW, this process appears to improve the overall quality of the returned results. However, as de-trending may not be appropriate for all datasets, we plan to implement this function as an optional step in the feature extraction process.

In the upcoming period of performance, we plan to complete implementation and evaluation of the Convergent Cross Mapping (CCM) technique, as a third, complementary technique for analyzing causality between time series.

Table 1: Automatic ranking of likely causal relationships between time series, using dynamic time warping.

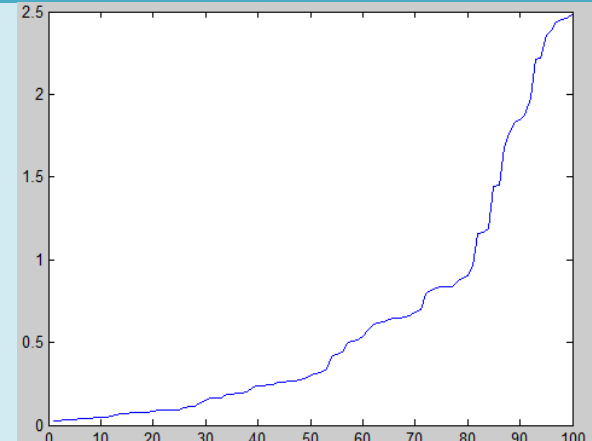
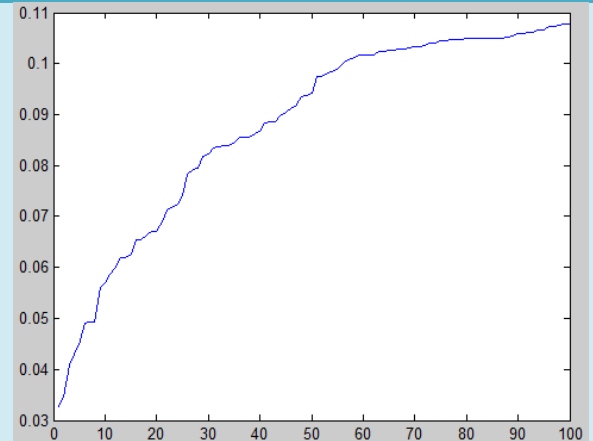
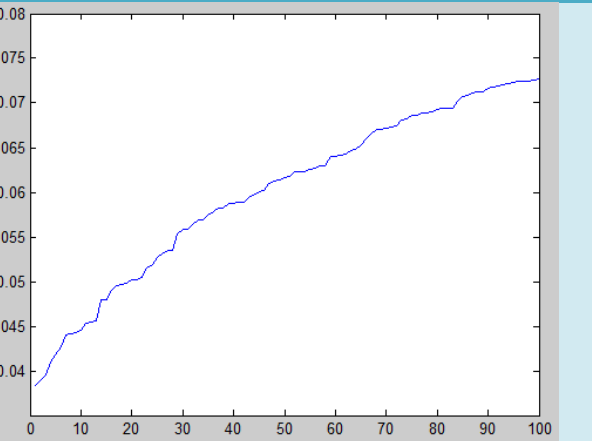
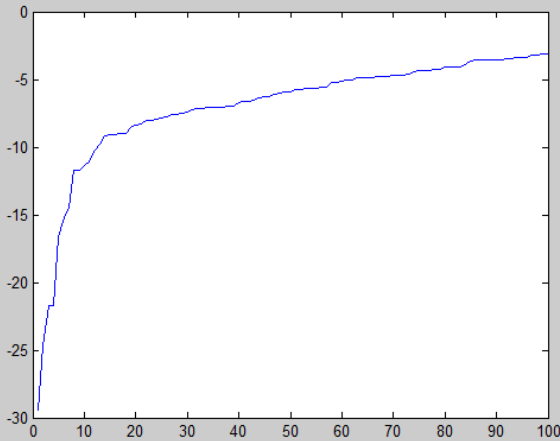
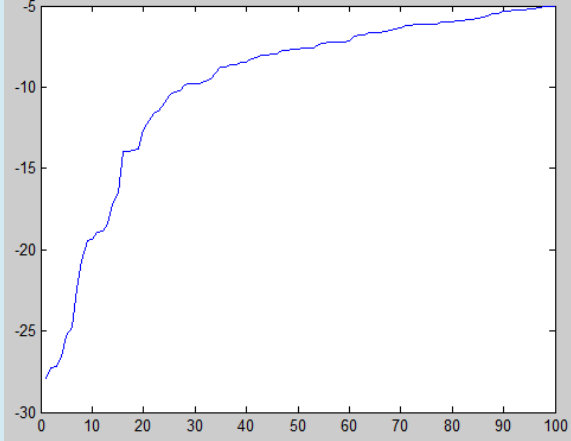
Natural Gas Rents: Top 20 Causally-Linked Series (Dynamic Time Warping)		
Raw Data	Scaled	Scaled and Detrended
'Arable land (hectares per person)' 'CO2 emissions (kg per 2000 US\$ of GDP)' 'CO2 emissions (kg per 2005 PPP \$ of GDP)' 'Adjusted savings: carbon dioxide damage (% of GNI)' 'Forest rents (% of GDP)' 'CO2 emissions (kg per PPP \$ of GDP)' 'Coal rents (% of GDP)' 'Mineral rents (% of GDP)' 'CO2 emissions from other sectors (% total fuel combustion)' 'Adjusted savings: mineral depletion (% of GNI)' 'Prevalence of HIV, total (% of population ages 15-49)' 'Permanent cropland (% of land area)' 'Adjusted savings: particulate emission damage (% of GNI)' 'Oil rents (% of GDP)' 'Foreign direct investment, net inflows (% of GDP)' 'DEC alternative conversion factor (LCU per US\$)' 'Official exchange rate (LCU per US\$, period average)' 'Foreign direct investment, net outflows (% of GDP)' 'Armed forces personnel (% of total labor force)' 'Personal transfers/compensations of empl. (% of GDP)'	'Total natural resources rents (% of GDP)' 'Electricity production from oil sources (% of total)' 'Adjusted savings: natural resources depletion (% of GNI)' 'Electricity production from oil sources (kWh)' 'Adjusted savings: energy depletion (% of GNI)' 'Inflation, consumer prices (annual %)' 'Real interest rate (%)' 'Inflation, GDP deflator (annual %)' 'Total reserves in months of imports' 'Oil rents (% of GDP)' 'Adjusted net savings, excl particulate emission damage (\$)' 'Mineral rents (% of GDP)' 'Claims on private sector (annl growth as % of broad money)' 'Adjusted savings: net national savings (% of GNI)' 'Adjusted savings: mineral depletion (% of GNI)' 'Adjusted savings: net national savings (current US\$)' 'Lending interest rate (%)' 'Coal rents (% of GDP)' 'Broad money growth (annual %)' 'Money and quasi money growth (annual %)'	'Total natural resources rents (% of GDP)' 'Adjusted savings: energy depletion (% of GNI)' 'Adjusted savings: natural resources depletion (% of GNI)' 'CO2 emissions from gaseous fuel consumption (kt) ' 'CO2 emissions (kg per 2000 US\$ of GDP)' 'Population in urban agglomerations of more than 1M (% total pop)' 'Forest rents (% of GDP)' 'Electricity production from oil sources (% of total)' 'CO2 emissions from liquid fuel consumption (% of total) ' 'CO2 emissions from residential bldgs/comm. + pub services (MMT) ' 'Fossil fuel energy consumption (% of total)' 'CO2 emissions from liquid fuel consumption (kt) ' 'Electricity production from oil sources (kWh)' 'Broad money to total reserves ratio' 'Money and quasi money (M2) to total reserves ratio' 'Quasi-liquid liabilities (% of GDP)' 'Electricity production from oil, gas and coal sources (% of total)' 'CO2 emissions from other sectors (MMT)' 'Industry, value added (% of GDP)' 'Oil rents (% of GDP)'
		

Table 2: Automatic ranking of likely causal relationships between time series, using granger causality.

Natural Gas Rents: Top 20 Causally-Linked Series (Granger Causality)		
Raw Data	Scaled	Scaled and Detrended
	'Road sector gasoline fuel consumption per capita (kg of oil eq)' 'CO2 emissions (metric tons per capita)' 'Electricity production from oil sources (kWh)' 'Energy use (kg of oil equivalent per capita)' 'Terms of trade adjustment (constant LCU)' 'Electricity production from oil sources (% of total)' 'Combustible renewables and waste (% of total energy)' 'CO2 emissions from liquid fuel consumption (% of total) ' 'CO2 emissions from liquid fuel consumption (kt) ' 'CO2 emissions from solid fuel consumption (% of total)' 'CO2 emissions from residential buildings and commercial)' 'Adjusted net savings, ex particulate emission damage' 'Coal rents (% of GDP)' 'CO2 emissions from manufacturing industries' 'Adjusted savings: net national savings (% of GNI)' 'Adjusted savings: consumption of fixed capital (% of GNI)' 'Land area (sq. km)' 'Surface area (sq. km)' 'Risk premium on lending (lending minus treasury rate)' 'Adjusted savings: energy depletion (current US\$)'	'Road sector energy consumption per capita (kg of oil eq)' 'Energy imports, net (% of energy use)' 'Road sector energy consumption (kt of oil equivalent)' 'Road sector gasoline fuel consumption (kt of oil equivalent)' 'Road sector gasoline fuel consumption per capita (kg of oil)' 'CO2 emissions from liquid fuel consumption (kt) ' 'Electricity production from oil sources (kWh)' 'Energy use (kt of oil equivalent)' 'CO2 emissions from transport (million metric tons)' 'Energy use (kg of oil equivalent per capita)' 'CO2 emissions (metric tons per capita)' 'Consumer price index (2005 = 100)' 'CO2 emissions (kt)' 'Electricity production from oil sources (% of total)' 'Stocks traded, turnover ratio (%)' 'Industry, value added (constant LCU)' 'Industry, value added (constant 2000 US\$)' 'CO2 emissions from liquid fuel consumption (% of total) ' 'CO2 emissions (kg per PPP \$ of GDP)' 'Personal transfers and compensation of employees (US\$)'
		

3.2. Automated Suggestions for Model Construction and Refinement

The Model Construction and Refinement capability in MAT is used to suggest changes to the causal model to improve its explanatory power. In this section, we describe our developing algorithms for identifying potentially interesting features in data, for identifying possible relationships between those features, and integrating this functionality into the MAT software and user interface.

3.2.1. Feature extraction

The feature extractor is used to identify potentially interesting “events” in temporal data. For instance, sudden dips, peaks, or times of instability are all possibly interesting events that can be used to reason about the data. The feature extractor looks for the constraint types already present in MAT code for the user-driven feature descriptions and associated feature learning:

- **constant**, in which the data remains within a particular range of some constant value;
- **slope**, in which the data increases or decreases at a particular rate;
- **spike**, in which the data increases sharply and then immediately decreases back to its original level or vice versa;
- **threshold**, in which the data crosses a particular boundary; and
- **train**, in which the data oscillates over a period of time.

The constraint generator’s goal is to generate features with high explanatory value. It is designed to generate features that are characteristic of a series, in that they appear more frequently in the series than in other series, but are not so common in the series that they have no explanatory value. The method used is similar to how many search engines perform keyword searches over a set of documents, except that the most common features are penalized.

For each constraint type, the constraint generator examines each data series and determines the number of times a feature corresponding to the constraint type appears in the data series as well as the density of the feature in the series, defined as the number of occurrences of the feature divided by the series’ length. The constraint generator then determines the density of the feature throughout the entire data set.

It then generates a score for each constraint using a variant of the tf-idf algorithm:

$$tfidf = tf \times \frac{-\log tf}{df}$$

where tf is “term frequency,” that is, the density of a feature in a given series, and df is “document frequency,” that is, the density of a feature throughout the data set. The $(-\log tf)$ term is added to penalize features that occur too often in a series and therefore do not have good explanatory value.

The constraint generator removes dominated constraints from the candidates. One constraint dominates another if it is conceptually narrower (that is, its minimums are at least as high and its maximums are at least as low) than the other and has at least the same tf-idf score. The constraints which remain are handed off to the feature extractor.

The feature extractor applies the constraints returned by the constraint generator to the data series, yielding a feature series for each data series in the data set.

3.2.2. User Interface

The new MAT functionality for suggesting changes to the model to improve its explanatory power is accessed through the Graph tab. After having selected a data set in the Visualization view and having added one or more nodes to the graph, the user may right-click a node in the Graph view and choose the “Recommend causes” option as shown in Figure 3.

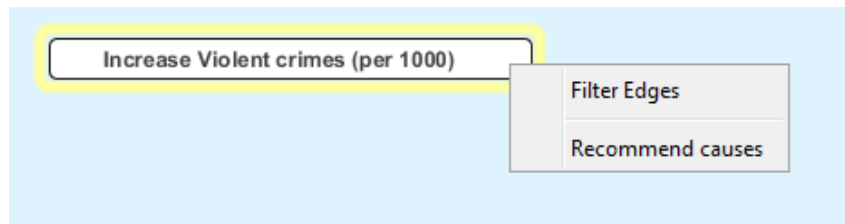


Figure 3. User choosing a concept in the model to find better explanations for

The user will then be presented with a dialog asking for the number of model-modification recommendations to generate. This is shown in Figure 4. Fewer recommendations may be presented if the requested number of recommendations cannot be found.

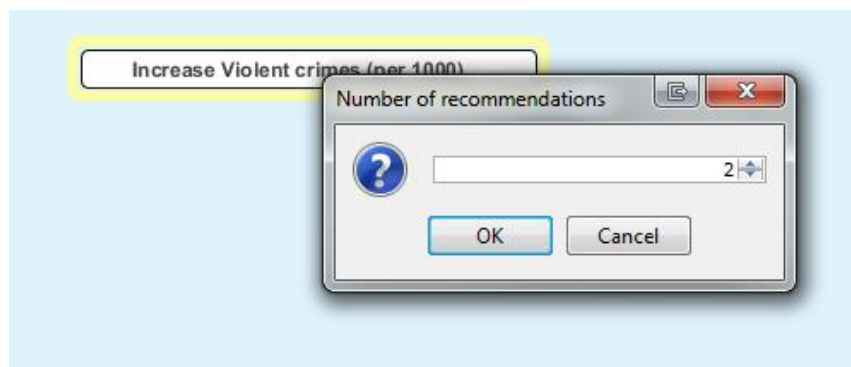


Figure 4. User specifying the number of model-refinement suggestions to generate

When the user presses OK, MAT will examine the data and attempt to find likely causes for the selected concept. The recommender begins by iterating over each data series in the same category as the selected concept; for the World Bank data set in the demo, the category is the country for which the data was collected. For each data series, the recommender generates features distinctive to the series, neither too common (and therefore likely noise) nor too rare (and therefore too easy to explain). The recommender is biased by design towards returning

features which are temporally narrow. This feature-discovery process is described in more detail in the previous report.

Once possibly interesting features have been discovered, the recommender then iterates over each feature to evaluate it as a potential cause, inserting it into a simple “A causes B” model and calculating the correlation score using the algorithm described in the previous report. The potential causes are ranked by correlation; potential causes with a score of zero are discarded. The recommender then plots the highest-ranking causes on the graph, up to the number specified by the user, as causes of the effect the user had selected. Figure 5 shows two possible explanations for *Increase in Violent Crimes* that MAT found by looking for correlated and, ideally, time-offset features in other available data streams.

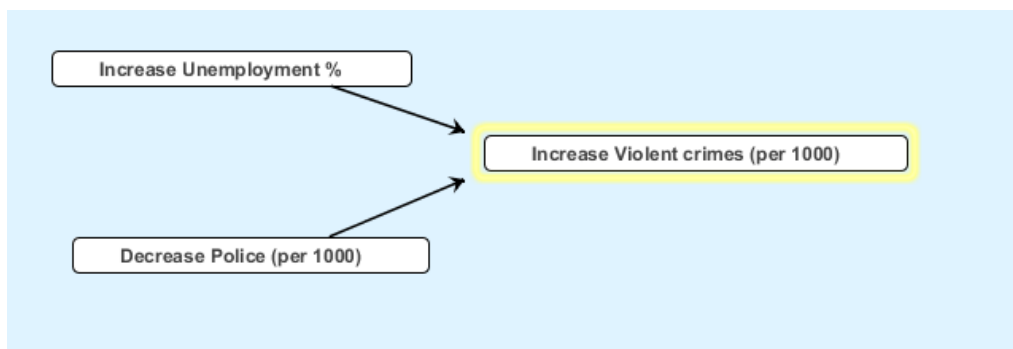


Figure 5. MAT suggests two possible explanations for increases in crime

The user may validate the resulting model in the Validate tab to see each candidate’s correlation value, both graphically and numerically (see Figure 6). Those suggestions that are not acceptable to the user can be deleted; those that are acceptable are left and can be saved as part of the improved model.

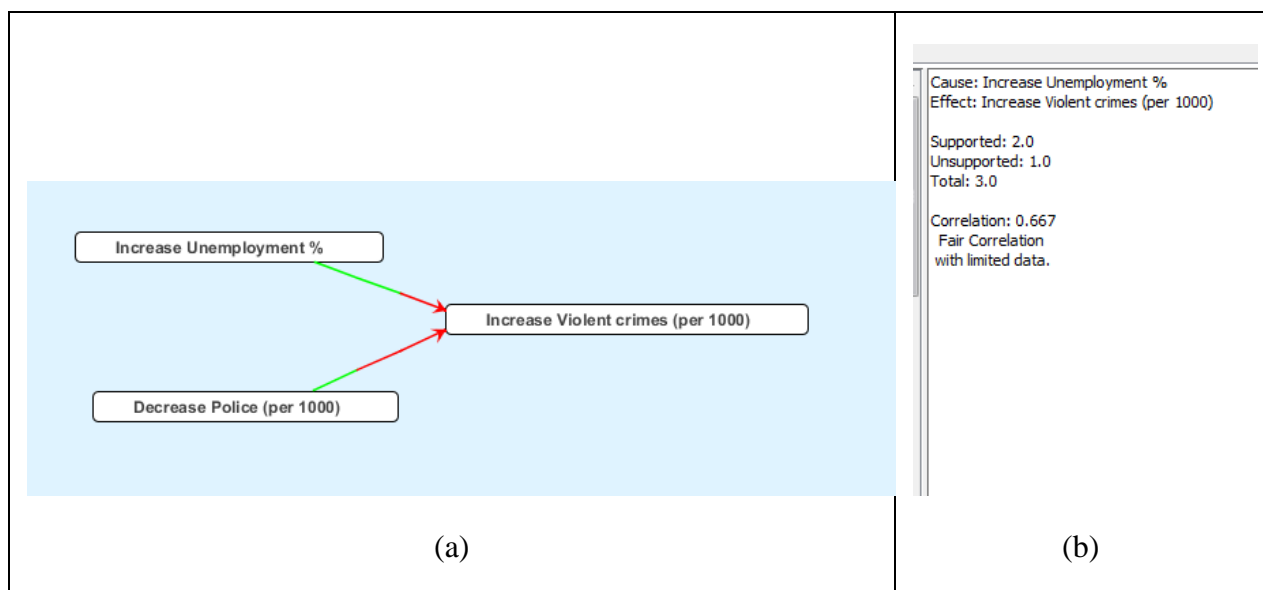


Figure 6. The user validating suggested explanations for crime in the validation tool, using both graphical and numerical methods

The user may also return to the Visualization tab to see the features that were generated for each candidate as shown in Figure 7.

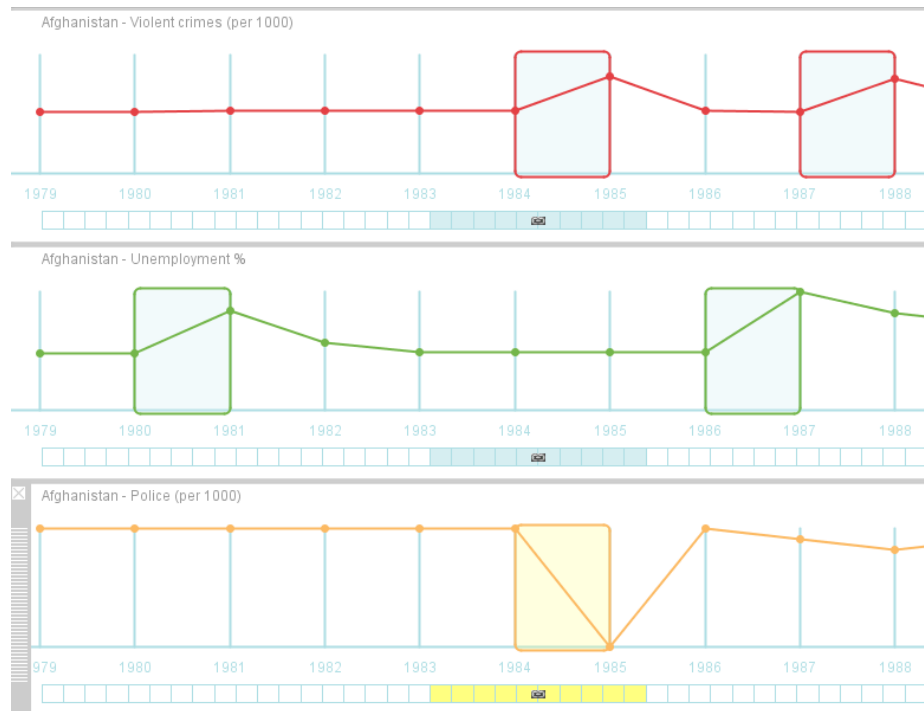


Figure 7. Features in unemployment and police force data that are discovered by MAT as potentially interesting and possibly related to increases in crime

3.3. Software Improvements

3.3.1. Improvements to Usability

One of our ongoing goals is to ensure that the MAT tool is actually useful to and usable by the scientific community. So, while most of this effort is focused on novel science, we are also constantly striving to ensure that the MAT software itself provides a usable framework for getting this science out to the community. To this end, we are currently working to improve the user interface for MAT with a focus on more customizability and better support for analyzing data and models across multiple types of analysis (such as by being able to view the validation and data analysis panes at the same time and to highlight and center on a feature in the validation pane when it is clicked on in the data visualization pane).

During the current reporting period, we focused on improving the graphical user interface by using Charles River's Metronome framework. This framework is built on top of the same Equinox libraries that the popular Eclipse Development Environment uses. In addition to increased robustness, the framework provides facilities for rearranging user interface components, thus providing the user with more flexibility when using MAT. For example, if

the names of the data series are long, then they could be cut off in the user interface, but after rearranging the components the names are fully visible (see Figure 8). The Metronome framework also provides functionality for undo and redo, so the user can easily correct mistakes.

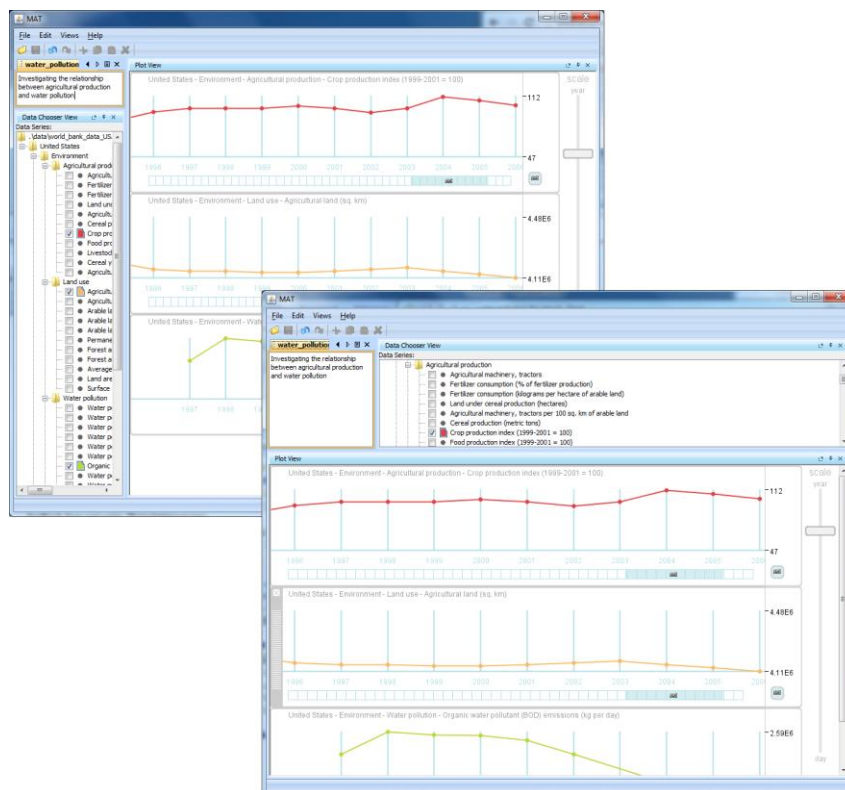


Figure 8. Changing Pane sizes and layouts in the new Metronome-enhanced MAT

This period, we also improved the MAT project file format so that changes in the user interface (e.g., color and layout of a data series) can be persisted once the MAT application has been closed.

3.3.2. New Functionality: Data Synthesis and Manipulation Capability

The ability to create new data, either de novo or, more typically, derived as a function of existing data series, gives the analyst the ability to explore complex relationships in the data and to envision hypothetical data that reflect imagined real-world possibilities. Consequently, we are extending MAT to enable it to synthesize new data series and pass them to the data visualization and validation components where they can be treated the same way as observed series. The addition of this new feature will enable users of the software to expand their corpus of data without additional collection and better process and handle signal-type data. As part of this new capability we will be adding the ability to make series-to-series data analytics in addition to the feature analytics and validation we are already supporting.

The suite of functionality for data synthesis will include a standard set of binary transforms on data (e.g., add one series to another) and the ability to define an arbitrary wave form using a Fourier series:

$$F(x) = a_0/2 + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x + \dots + a_n \cos nx + b_n \sin nx + \dots$$

The new toolkit will also allow the user to break a series into its constituent fundamental waves using standard Fourier analysis (FFT). The binary transforms to combine two series into a resultant synthetic series are as follows:

moving average: $y = \text{avg}(y'[-n], \dots, y'[-2], y'[-1], y', y'[+1], y'[+2], \dots, y'[+n])$

sum/difference between two series: $y = y1' \pm y2'$

series product: $y = y1' * y2'$

linear convolution: $y = y1'm1 + y2'm2 + b$

We also have a set of mono transforms that operate on a single series to produce a resultant:

linear scaling: $y = y'm + b$

logarithmic scaling: $y = \log(y')$

inversion: $y = 1 / y'$

The interface for the new tools reuses the existing Graphler module in MAT which is used for creating graphical models. Graphler has been improved so that it can now support different types of graphical editing and is no longer hard-coded to produce feature models alone. The design concept for the UI is shown in Figure 9. New tool icons are being added on the palette; these will enable operations to be performed on data series analogously to the way logical operations are currently performed on features in models. The data properties pane is being repurposed to allow the user to define the parameters of the transformation. The node list in this view shows a list of all data series, both real and synthetic, instead of feature concepts.

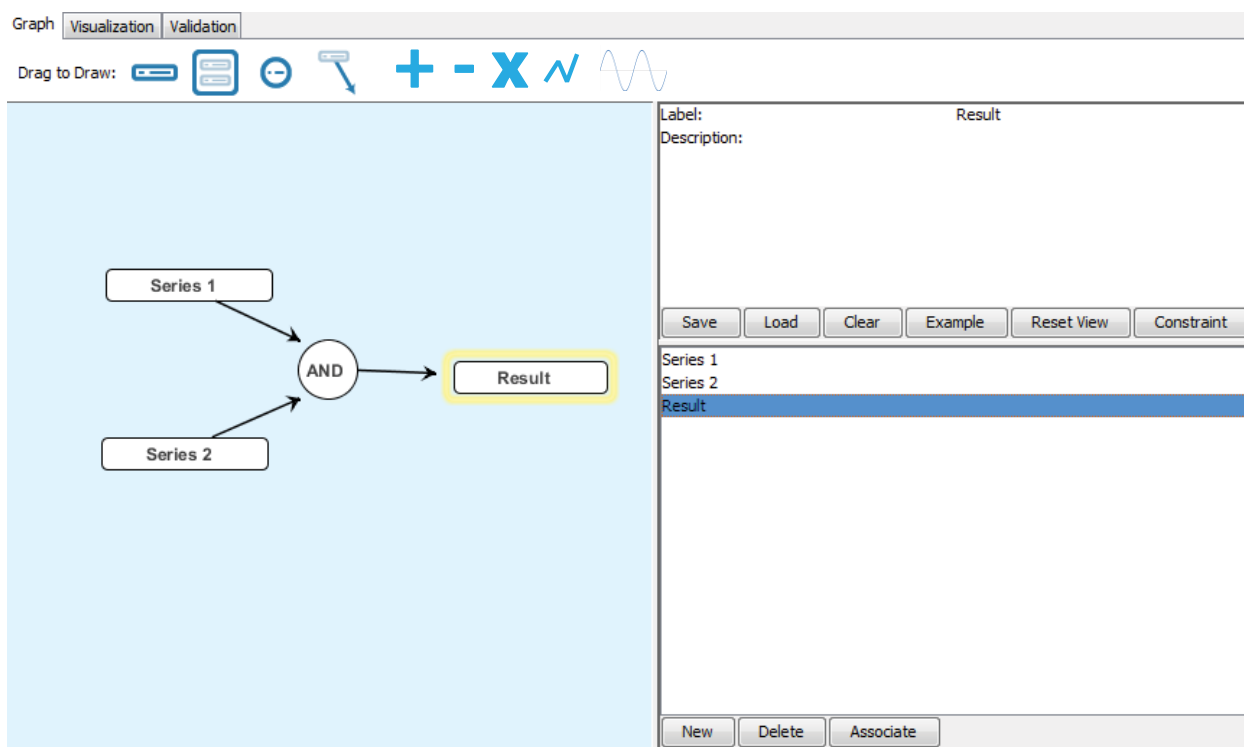


Figure 9. Creating a new temporal data series in MAT

4. Planned Activities

The incorporation of DTW into MAT is in its infancy and there is still work to be done on improving the visual aspect as well as core functionality to make this tool beneficial to the user. These include items such as:

- Provide the user the ability to limit how much warping is permissible
- Display patterns in the warping such as shift, compression, and expansion regions
- Incorporate DTW into the recommendation tool for uncovering relationships
- Deeper level of integration of the DTW libraries to the MAT code to provide speed improvements

During the upcoming period, we plan port our Matlab implementations of GC and CCM to Java and integrate them into the MAT tool

We will also continue to improve our recommendation system. Some of the improvements we are planning for the upcoming period include:

- Considering time lag as a weighting factor in scoring, with closer temporal proximity between cause and effect yielding proportionally higher scores
- Integrating dynamic time warping to consider varying times between cause and effect

- Increasing the complexity of the causal model used by the scoring module, considering multiple causes for a given effect and multiple levels of causality
- Using machine learning techniques to consider multiple potential designs for causality models

We will also continue to explore applications of MAT to data and problems in neurophysiological modeling.

Finally, we will begin work on data validation, which is the final research area for this effort to be addressed.

5. Budget and Project Tracking

As of April 30, 2013, we have spent \$224,851, or 24% of our total budget of \$928,224, in 24% of the scheduled time. Our current funding is \$362,445, so we have spent 62% of our available funding.

We believe we are in good shape to complete the project on time and on budget.

6. References

- Chan, P. & Salvador, S. (2007). Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intelligent Data Analysis*, 11(5).
- Dereszynski, E. & Dietterich, T. (2007). Probabilistic Models for Anomaly Detection in Remote Sensor Data Streams.
- Dereszynski, E. W. & Dietterich, T. G. (2011). Spatiotemporal Models for Data-Anomaly Detection in Dynamic Environmental Monitoring Campaigns. *ACM Transactions on Sensor Networks (TOSN)*, 8(1).
- Neal Reilly, S. (2010). *Validation Coverage Toolkit for HSCB Models*. (Rep. No. R09005-03). Cambridge, MA: Charles River Analytics Inc.
- Neal Reilly, W. S., Pfeffer, A., & Barnett, J. (2010). A Metamodel Description Language for HSCB Modeling. In D. Schmorow & D. Nicholson (Eds.), *Advances in Cross-Cultural Decision Making*. CRC Press.
- Neal Reilly, W. S., Pfeffer, A., Barnett, J., Chamberlain, J., & Casstevens, R. (2011). A Computational Toolset for Socio-Cultural Data Exploration, Model Refinement, and Model Validation. In *Proceedings of Human Social Culture Behavior (HSCB) Focus*. Chantilly, VA.